ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Predictions of Colloidal Molecular Aggregation Using AI/ML Models

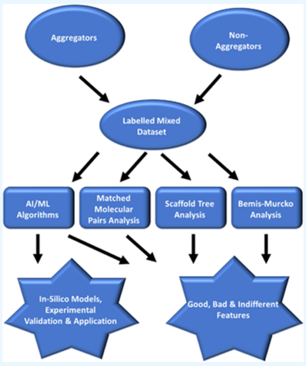
David C. Kombo,* J. David Stepp, Sungtaek Lim, Bettina Elshorst, Yi Li, Laura Cato, Maysoun Shomali, David Fink, and Matthew J. LaMarche

Cite This: *ACS Omega* 2024, 9, 28691−28706

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: To facilitate the triage of hits from small molecule screens, we have used various AI/ML techniques and experimentally observed data sets to build models aimed at predicting colloidal aggregation of small organic molecules in aqueous solution. We have found that Naïve Bayesian and deep neural networks outperform logistic regression, recursive partitioning tree, support vector machine, and random forest techniques by having the lowest balanced error rate (BER) for the test set. Derived predictive classification models consistently and successfully discriminated aggregator molecules from nonaggregator hits. An analysis of molecular descriptors in favor of colloidal aggregation confirms previous observations (hydrophobicity, molecular weight, and solubility) in addition to undescribed molecular descriptors such as the fraction of sp^3 carbon atoms (Fsp3), and electrotopological state of hydroxyl groups (ES_Sum_sOH). Naïve Bayesian modeling and scaffold tree analysis have revealed chemical features/scaffolds contributing the most to colloidal aggregation and nonaggregation, respectively. These results highlight the importance of scaffolds with high Fsp3 values in promoting nonaggregation. Matched molecular pair analysis (MMPA) has also deciphered context-dependent substitutions, which can be used to design nonaggregator molecules. We found that most matched molecular pairs have a neutral effect on aggregation propensity. We have prospectively applied our predictive models to assist in chemical library triage for optimal plate selection diversity and purchase for high throughput screening (HTS) in drug discovery projects.


Aggregators
Non-Aggregators
Labelled Mixed Dataset
AI/ML Algorithms
Matched Molecular Pairs Analysis
Scaffold Tree Analysis
Bemis-Murcko Analysis
In-Silico Models, Experimental Validation & Application
Good, Bad & Indifferent Features

1. INTRODUCTION

It is well known that at micromolar and submicromolar concentrations, various organic molecules, including some marketed drugs, can form colloidal aggregates.[1−3] Such a state, which is intermediate between true solution and precipitate, can result in nonspecific assay effects through sequestration, denaturation, or conformational change of the target protein. Colloidal aggregation is, unfortunately, and undoubtedly a real challenge to drug discovery and development. A more thorough understanding of the interplay between various physicochemical properties and molecular features akin to colloidal aggregation is important not only for molecular design and hit optimization but also for measuring bioactivity in cell culture, drug formulation, drug delivery, triage of HTS hits, the purchase of external chemical libraries to enrich compound collections, and the design of DNA-encoded libraries.

In response to the Covid-19 pandemic, drug repositioning has been widely used to search for antiviral drugs with therapeutic value for SARS CoV-2 targets.[4−8] Recently, O'Donnell et al.[9] have investigated the role of colloidal aggregation as a source of false positives and artifacts in drug repurposing studies. They found that some of the hits identified through biochemical assays can act as colloidal aggregators at the concentrations used for screening, typically in the micromolar range. This result clearly suggests that an understanding of molecular properties and chemical features driving colloidal aggregation may help in preselecting a more appropriate chemical library for interrogation. Thus, we propose that AI/ML models might be useful to proactively predict the propensity of molecules to form colloidal aggregates, thereby reducing the number of screening artifacts and nonspecific effects.

During the last two decades, numerous computational studies aimed at predicting colloidal aggregation have been reported. Seidler et al.[10] used Cerius2[11] descriptors to build a recursive partitioning (RP) model from a training set of 111 compounds, out of which 47 were aggregators and 64 were not. This model scored very well on the training set, with an accuracy of 94% but with a reduced accuracy of 79% on the test set of 75 drug molecules. Feng et al.[12] subsequently built RP, random forest (RF), and Naive Bayesian (NB) models from a training set of 732 compounds. Once validated against a random test set of 298 compounds, these models predicted aggregation at a true positive rate of 77% (RP), 60% (RF), 23% (initial NB), and 74% (refined NB). Corresponding misclassification rates for these models were 26% (RP), 11% (RF), 26% (initial NB), and 20% (refined NB). Rao et al.[13] used a data set comprised of 1319 aggregators and 128,325

Received: March 28, 2024
Revised: June 10, 2024
Accepted: June 12, 2024
Published: June 18, 2024

ACS Publications

https://doi.org/10.1021/acsomega.4c02886

nonaggregators to build support vector machine (SVM), K-nearest neighbors (KNN), and continuous kernel discrimination (CKD) models aimed at classifying aggregators and obtained a 5-fold cross-validation-derived sensitivity (also referred to as true positive rate or recall) of 77.8, 71.6, and 74.7%, respectively. Irwin et al.[14] used 2D chemical similarity to known aggregators and lipophilicity ($C\log P > 3$) to predict colloidal aggregation. They prospectively found that a total of 18 molecules with Tanimoto coefficients between 0.85 and 0.99 to known aggregators indeed aggregated at relevant concentrations. Yang et al.[15] have recently used a data set comprised of 12,119 aggregators and 24,172 nonaggregators to develop RP, RF, and extreme gradient boosting (XGBoost) decision tree models with 5 sets of groups of molecular descriptors. The best-derived model had an accuracy of 0.95 and 0.94 for the training and test sets, respectively. A recently published review article discussed experimental methods for the detection of colloidal aggregates and various computational approaches, which provided additional information and insight.[16]

Although several computational efforts have focused on building predictive models for colloidal aggregation, there has been no report to decipher matched molecular pairs between aggregator and nonaggregator molecules. Enumeration of such matched pairs could be beneficial for the design of small molecules outside of the aggregator chemical space. It can also help in hit and lead optimization to remove aggregation features and include more desirable drug-like[17−19] chemical features. Moreover, many molecular descriptors related to drug-likeness and compound developability have not been used in previous modeling studies on colloidal aggregation. Examples include the fraction of carbon atoms sp^3 (Fsp3),[20] the number of stereocenters,[20] quantitative estimate of drug-likeness (QED),[21] and promiscuity-related descriptors such as stereochemical complexity and shape complexity.[22−25]

In the present study, we have used AI/ML techniques such as deep neural networks and logistic regression, which have not yet been utilized so far to predict colloidal aggregation, coupled with unexplored molecular descriptors relevant to reducing drug attrition. We have built models with excellent performance metrics such as accuracy, precision, recall, and BER on both the training and test sets. We have identified the importance of molecular descriptors, chemical features, and scaffolds promoting either aggregation or nonaggregation, using various analytical approaches. Our results outperform and are consistent with previous observations but also add new insights. We also report our findings on molecular matched pairs analysis, which can be applied in a context-dependent manner. Application of the derived models in HTS hits triage and compound acquisition has also been performed and reported herein.

2. METHODS

2.1. Computational Model Building and Validation.

Exploratory data analysis and predictive analytics were performed solely using BIOVIA Pipeline Pilot.[26] The data set was made of colloidal aggregators retrieved from the ZINC Database,[27] and from a published data set.[15] Nonaggregator molecules were retrieved from the Drug Central Database[28] and the experimental drugs were retrieved from TTD (Therapeutic Target Database).[29] After filtering and removing duplicate molecules, inorganic compounds, drugs known to be aggregators,[9] and those with ambiguous behavior, the total number of the remaining compounds was 29,345. Training and test sets were randomly selected to contain 70 and 30% of compounds, respectively, and the total number of compounds in the training and test sets were 20,572 and 8773, respectively.

MMPA was carried out within Pipeline Pilot, using one cut, a minimum core size of 5, and a maximum fragment size of 10 heavy atoms. Colloidal aggregators and nonaggregator molecules were respectively labeled 1 and 0. To build classification models, the following molecular descriptors as implemented within Pipeline Pilot, were used as independent variables: molecular weight (MW), $A\log P$, QED, Fsp3, number of atoms, number of rings, number of aromatic rings, number of rotatable bonds, number of hydrogen-bond acceptors (HBAs), number of hydrogen-bond donors (HBDs), (nitrogen count + oxygen count), molecular solubility, polar surface area (PSA), molecular surface area, fractional PSA, solvent-accessibility surface area, extended connectivity functional class molecular fingerprints, up to six bonds (ECFP-6), chi indices, Balaban, Wiener and Zagreb indices, Kappa shape indices, subgraph counts, and electrotopological state keys. In addition to using synthetic accessibility score (SAscore) as an estimate of molecular complexity,[30] we have used the number of stereocenters, stereochemical complexity, and shape complexity, defined as follows:

$$\text{Num of stereo centers} = (\text{NumUnknownTrueStereoAtoms} + \text{NumTrueStereoAtoms}) \quad (1)$$

where Num designates the number.

$$\text{stereo chemical complexity} = (C_{\text{stereogenic}}/C_{\text{total}}) \quad (2)$$

where $C_{\text{stereogenic}}$ and C_{total} denote the number of C atoms that are stereogenic and the total number of C atoms, respectively. To calculate the parameter $C_{\text{stereogenic}}$, we have used eq 1.

$$\text{shape complexity} = (C_{sp3}/[C_{sp3} + C_{sp2}] \quad (3)$$

where C_{sp3} and C_{sp2} denote the number of sp^3- and sp^2-hybridized carbons atoms, respectively.

Classification models were built using Naïve Bayesian, random forest, recursive partitioning tree, support vector machine, logistic regression, and deep neural network, based on the above-mentioned molecular descriptors. All models were built in Pipeline Pilot as well. For random forest models, 500 trees were used to build the ensemble. The number of descriptors used for each tree in the forest was equal to the square root of the total number of descriptors. To reduce bias due to data set imbalance and to attempt to increase both accuracy and speed, class sizes were equalized by creating a random data set containing an approximately equal number of aggregators and nonaggregators for each tree in the forest. For the deep neural network, we used a sigmoidal activation function for nodes in the network, 2 hidden layers, and 50 nodes per hidden layer. The maximum number of iterations for network training was 5000. Momentum and learning rates were 0.9 and 0.05, respectively. Both dropout fractions for hidden layers and visible layers were 0.25. For logistic regression model building, we used the glm (generalized linear model) method in R, which performs a logistic regression without bias correction. We also built two-class Bayesian categorization models aimed at distinguishing nonaggregator data records from aggregator records, using scaled probabilities

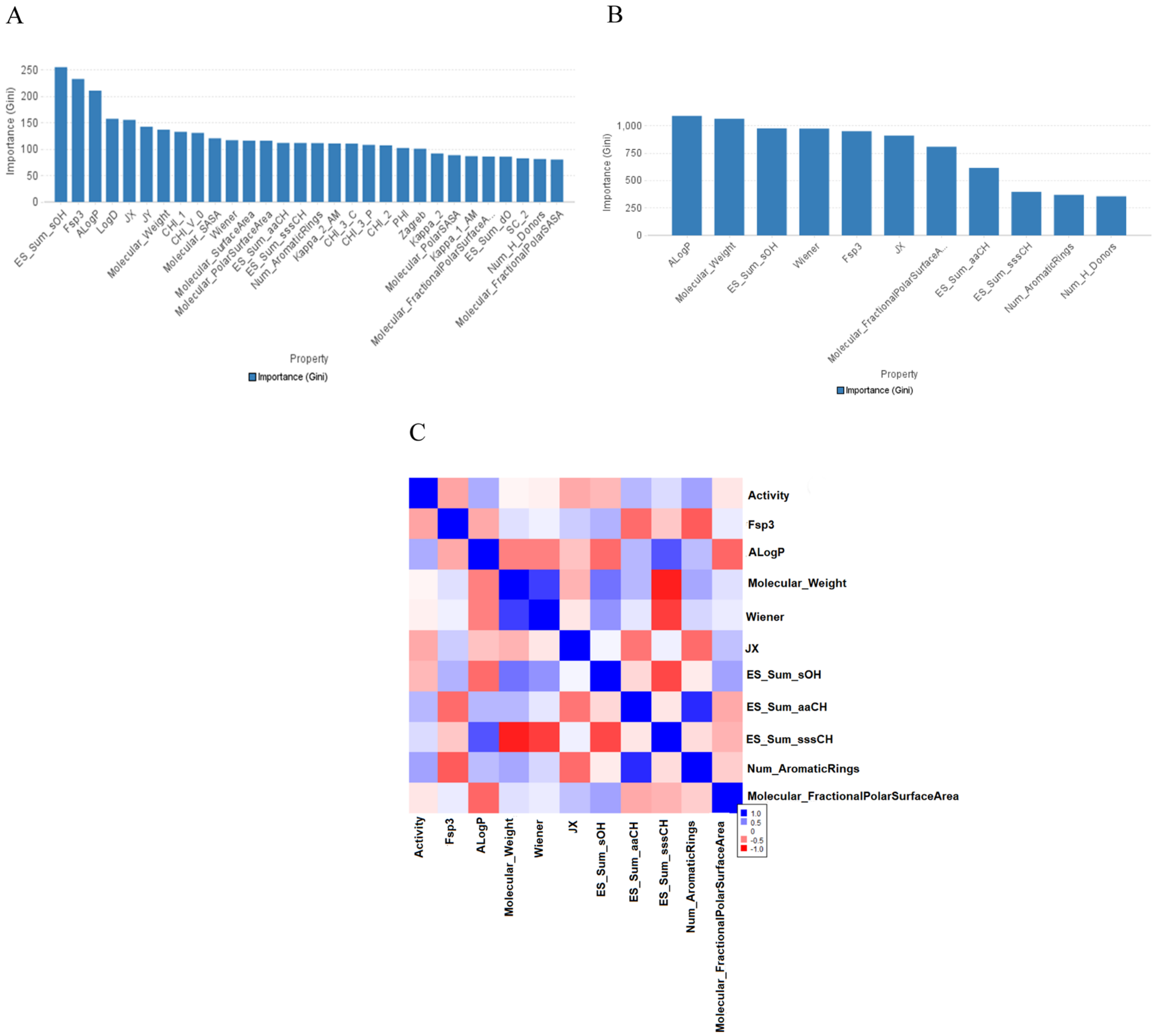


Figure 1. Molecular descriptors important for colloidal aggregation, as derived from the random forest classification model. (A) Only the top-30 descriptors are shown. (B) After removal of highly correlated descriptors. (C) Pearson correlation heatmap of the top important descriptors shown in panel (B).

derived from these record counts. The recursive partitioning tree model was built using the following parameters: Maximum tree depth:50, Gini as the split method, a weighting by class, maximum knots per property equal to 50, both maximum generic depth and maximum lookahead depth equal to 0, and a minimum yes or no answers equal to 1. For the support vector machine model, we used a C-classification type, a radial kernel, a weighting method as uniform, and episillon equal to 0.1, and the cost associated with training set error was set to 2.0.

To compare various models, the following performance metrics were used: Receiver operating characteristic curve (ROC AUC), accuracy, precision, recall, *F*1 score, and misclassification error, which are derived from true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as follows:

ROC AUC = Area under the curve derived by plotting the true positives rate vs the false positive rate.

$$\text{accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{4}$$

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{5}$$

$$\text{recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{6}$$

$$F1\text{score} = 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall}) \tag{7}$$

$$\begin{aligned}&\text{balanced error rate(BER)}\\&\quad = 0.5 \times (\text{FP}/(\text{TN} + \text{FP}) + \text{FN}/(\text{FN} + \text{TP}))\end{aligned} \tag{8}$$

$$\text{misclassification error(aggregators)}: \text{FN}/(\text{TP} + \text{FN}) \tag{9}$$

$$\text{misclassification error(nonaggregators)}: \text{FP}/(\text{TN} + \text{FP}) \tag{10}$$

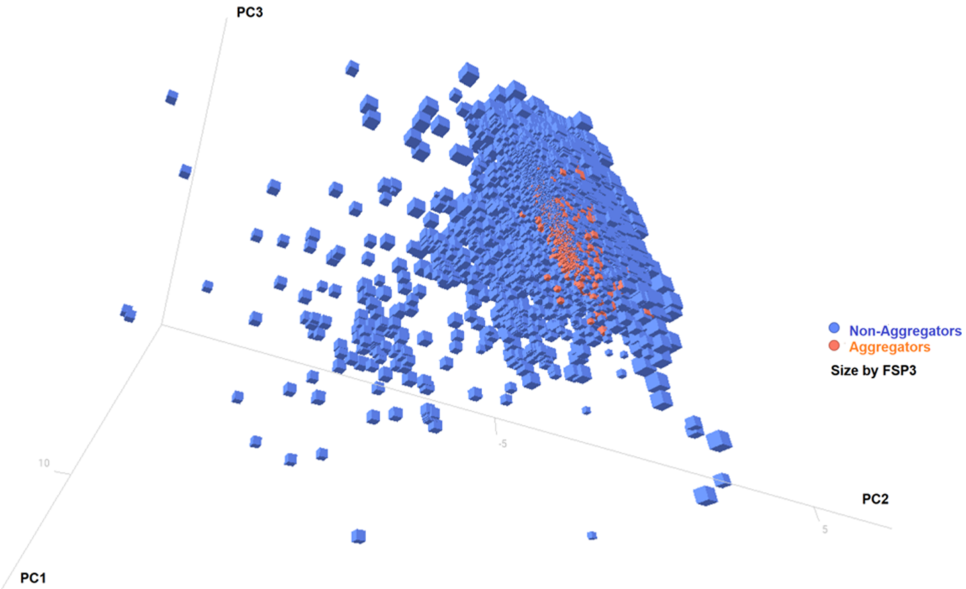


Figure 2. 3D plots of PCA components for the whole data set. Aggregators and nonaggregator molecules are colored red and blue, respectively. Data points are shown as filled cubes. The larger the latter, the higher the Fsp3 value of the compound.

$$\text{misclassification rate(MR)}: (\text{FN} + \text{FP})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (11)$$

2.2. Experimental Measurements by NMR WaterLOGSY.

Ninety-nine in-house compounds were tested experimentally for the formation of colloidal aggregation using NMR WaterLOGSY[31] and saturation transfer difference (STD) measurements. The NMR buffer was composed of 10 mM Na_2HPO_4, 154 mM NaCl, and 5% D_2O, pH 7.4. The concentration of the NMR samples was 200 μM compound in buffer with 100 μM TSPA (internal reference), with a 160 μL volume per 3 mm NMR tubes. NMR measurements were performed using a Bruker 700 MHz spectrometer at 298 K equipped with a cryoprobe. For each sample, a ^{1}H 1D spectrum (zgesgp, expt 5 min, NS 128), a WaterLOGSY spectrum (ephogsypgno.2, expt 38 min, NS 512), and a STD spectrum (stddiffesgp.3, expt 48 min, NS 512) were recorded. For nonaggregating compounds with sufficient solubility in aqueous solution, negative signals were detected in the WaterLOGSY spectra and no signal was found in the STD spectrum. However, if the compound forms colloidal aggregates or micelles in the aqueous solution, in the WaterLOGSY spectrum, no signal or even positive signals were detected. In the corresponding STD spectrum, positive signals were measured. These results indicate that the average molecular weight of the compounds is significantly higher than the monomeric, solubilized form, and therefore, the compound forms aggregates or micelles in aqueous solution. Compounds that were found insoluble and for which no NMR signal was detected in the ^{1}H 1D spectrum were not included in the validation set. For 14 compounds, aggregation or micelle formation was detected due to positive signals in the WaterLOGSY and STD spectrum. The remaining 85 compounds were soluble and monomeric in solution since the WaterLOGSY spectra showed negative signals.

3. RESULTS

3.1. Molecular Descriptors Important for Colloidal Aggregation.

We have applied random forest to decipher among the large pool of descriptors described in Section 2 those important in discriminating colloidal aggregators from molecules tending not to aggregate. Using all the descriptors, we obtained a predictive model with a ROC score (out-of-bag data) of 0.96 and a misclassification error rate of 10.3%. The corresponding results for the test set were a ROC score of 0.89 and a balanced error rate of 11%. Results shown in Figure 1A suggest that the following descriptors, as listed in decreasing order of importance, contribute to colloidal aggregation: ES_Sum_sOH, Fsp3, AlogP, LogD, Molecular_Weight, JY, JX, Molecular_SASA, CHI_V_0, number of aromatic rings, ES_Sum_sssCH, CHI_1, CHI_3_P, Molecular_PolarSurfaceArea, CHI_3_C, molecular surface area, CHI_2, ES_Sum_aaCH, Wiener, Kappa_2, Kappa_1_AM, Molecular_PolarSASA, Kappa_2_AM, SC_0, CHI_0, ES_Sum_dO, Molecular_FractionalPolarSurfaceArea, Kappa_3_AM, Zagreb, and PHI. To avoid overfitting, we selected descriptors with an absolute value of Pearson correlation coefficient of <0.90. In the cases of pairs of highly correlated descriptors with the absolute value of Pearson correlation coefficient ≥0.90, we only retained the independent variable of the pair with the highest contribution to the model as shown in Figure 1A. The derived final list of independent variables used in the subsequent model building was composed of the following: Fsp3, ALogP, molecular weight, number of H-bond donors, number of aromatic rings, Wiener index, ES_Sum_sOH, ES_Sum_sssCH, ES_Sum_aaCH, molecular fractional polar surface area, and JX. Their respective feature importance to the random forest model is shown in Figure 1B, while their correlation matrix heatmap is shown in Figure 1C. In another scenario, from the pairs of highly correlated descriptors with the absolute value of Pearson correlation coefficient ≥0.90, we retained the independent variable with the highest correlation with the colloidal aggregation activity. This second set of features was composed of the following molecular descriptors: Fsp3, $A\log P$, PHI, number of aromatic rings, Wiener index, ES_Sum_sOH, ES_Sum_sssCH, ES_Sum_aaCH, molecular fractional polar solvent-accessibility surface area, and JX. Models derived using either set of descriptors were compared via their performance metrics and the best-scoring models were selected for further applications, as described below.

To further probe features that are important for aggregation, we reduced the dimensional space of our molecular descriptor sets by carrying out a principal component analysis. We found that 3 components can cumulatively explain 98% of the variance of our data set via a linear combination of the following 4 molecular descriptors: Fsp3, molecular weight, molecular solubility, and AlogP. This result highlights the importance of Fsp3 and is consistent with previous studies that showed the importance of molecular weight, hydrophobicity, and solubility in discriminating aggregators.[12−15] Figure 2 shows the PCA 3D plot which visualizes a grouping of aggregators and nonaggregators into 2 clusters.

3.2. Performance Metrics of the Models. To confirm whether our data sets could yield good predictive models with a correct classification rate above 0.70, we have calculated the modelability index (MODI).[32] For binary classification data, Golbraik et al. have demonstrated that a MODI ≥ 0.65 indicates that a data set is modelable. As shown in Table 1, we found a MODI of 0.84, when we used descriptors identified as both noncorrelating and important for aggregation classification. We also found that adding ECFP-6 descriptors to this set of important descriptors can improve the MODI to 0.92. Even when ECFP-6 was the only descriptor used to build models, the derived MODI of 0.92 is still high enough. Our finding that these MODI values are well above the expected threshold value of 0.65 suggests that our data sets will produce highly performant predictive models.

Table 1. Modelability Index Derived from the Training Set and Molecular Descriptors Used for Model Building

molecular descriptors	MODI	MODI_term (nonaggregators)	MODI_term (aggregators)
ECFP-6	0.92	0.94	0.90
important descriptors	0.84	0.86	0.82
ECFP-6 and important descriptors	0.92	0.92	0.91

Molecular descriptors that were found to be important to classify colloidal aggregation as previously described, and ECFP-6 fingerprints which yielded an excellent modelability index were used to build models. As shown in Table 2, all derived models are highly performant on the training set, as judged by the following performance metrics: accuracy, roc score, precision, recall, and *F*1 score, which are in the ranges (0.86−0.99), (0.94−0.99), (0.88−0.99), (0.88−0.99), and (0.88−0.99), respectively. Likewise, as shown in Table 3, all derived models are highly performant on the test set, as revealed by the following metrics: accuracy, roc score, precision, recall, and F1 score, which are in the ranges (0.83−0.90), (0.82−0.98), (0.86−0.94), (0.86−0.94), and (0.86−0.94), respectively. We find that the recursive partitioning tree, deep neural network, and Naïve Bayesian outperform the other models for the training set, achieving the lowest values of BER of 0.005, 0.04, and 0.07, respectively. We also found that the Naïve Bayesian model followed by the deep neural network model outperformed the other models for the test set, achieving the lowest BER values of 0.08 and 0.10, respectively. We then used a validation set consisting of an internal data set of 99 compounds experimentally tested for colloidal aggregation as described in the Section 2. Examples of structures investigated in this study are listed in Figure 3. Results shown in Table 4 indicate that all derived models are still reasonably performant on the validation set, as judged by the following performance metrics: accuracy, roc score, precision, recall, and *F*1 score, which are in the ranges (0.73−0.83), (0.57−0.76), (0.88−0.95), (0.79−0.86), and (0.83−0.90), respectively. Although the precision of the models is in the same ballpark for all three sets, the area under the ROC curve and BER of the validation set exhibit the lowest values (in the range 0.57−0.76) and the highest values (in the range 0.24−0.43), respectively.

Table 3. Test Set Summary Statistics and Performance Metrics of the Models

model	accuracy	ROC score	precision	recall	*F*1 score	BER
random forest	0.87	0.95	0.91	0.88	0.89	0.13
logistic regression	0.85	0.85	0.88	0.87	0.88	0.15
Bayesian	0.93	0.98	0.94	0.94	0.94	0.08
deep neural network	0.90	0.96	0.93	0.90	0.91	0.10
recursive partitioning tree	0.83	0.82	0.86	0.86	0.86	0.18
support vector machine	0.86	0.93	0.88	0.87	0.88	0.15

3.3. Structural Features Important for Colloidal Aggregation. In the previous computational studies, investigations aimed at deciphering structural features responsible for colloidal aggregation were only reported for Bemis-Murcko framework analysis.[36] In the current study, molecular features responsible for colloidal aggregation were derived from a Bayesian categorization model built by using ECFP-6 descriptors. Results shown in Figure 4A,4B indicate that nonaggregators appear to be richer in sp3 carbon atoms and poorer in the number of aromatic rings than aggregators. The latter class of molecules appears to be rich in sulfur-containing chemical moieties and aromatic heterocyclic groups, as previously observed by Yang et al.[15] Out of the top-20 features that contribute to colloidal aggregation, we find the following 12 features: B2, B3, B4, B8, B10, B11, B12, B13, B14, B16, B17 and B18 contain at least one sulfur atom, as shown in Figure 4B.

Table 2. Training Set Summary Statistics and Performance Metrics of the Models

model	accuracy	cross-validated ROC score	precision	recall	*F*1 score	BER
random forest	0.88	0.96	0.91	0.88	0.90	0.12
logistic regression	0.86	0.94	0.88	0.89	0.88	0.14
Bayesian	0.93	0.98	0.94	0.94	0.94	0.07
deep neural network	0.96	0.99	0.97	0.95	0.96	0.04
recursive partitioning tree	0.99	0.99	0.99	0.99	0.99	0.005
support vector machine	0.87	0.93	0.88	0.89	0.89	0.14

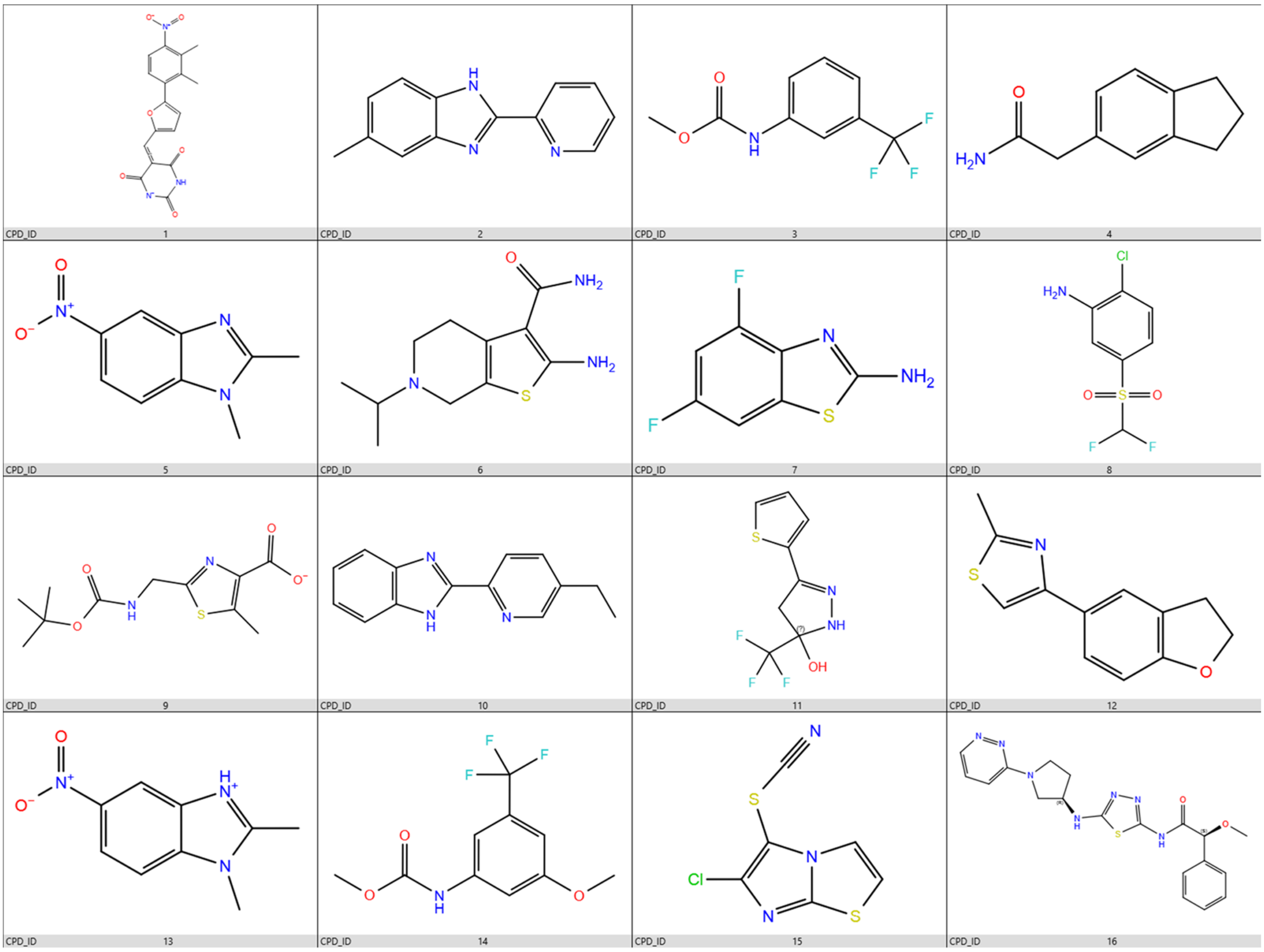


Figure 3. Examples of compounds that were part of the validation set. These compounds were purchased from Mcule,[33] Emolecules,[34] and Chembridge.[35] Only compound 7 was an experimentally observed aggregator. All the remaining compounds were nonaggregators.

Table 4. Validation Set Summary Statistics and Performance Metrics of the Models

model	accuracy	ROC score	precision	recall	*F*1 score	BER
random forest	0.80	0.76	0.95	0.81	0.87	0.24
logistic regression	0.81	0.71	0.92	0.85	0.88	0.29
Bayesian	0.83	0.75	0.94	0.86	0.90	0.25
deep neural network	0.77	0.70	0.91	0.81	0.86	0.34
recursive partitioning tree	0.73	0.57	0.88	0.79	0.83	0.43
support vector machine	0.77	0.69	0.92	0.80	0.86	0.31

To further investigate chemical features and scaffolds akin to colloidal aggregation, we generated scaffold trees for the whole data set. Input molecules were first trimmed to derive the Burma's−Murcko framework. Next, hierarchical ring scaffolds were derived using the iterative ring trimming procedure which yielded the hierarchical scaffold tree.[37] Histograms visualizing the distribution of the number of levels in the scaffold trees are significantly different between the aggregator and non-aggregator classes of molecules (Figure 5A,B). While the number of levels in the scaffold tree appears to be normally distributed for the colloidal aggregators, the corresponding histogram obtained for the nonaggregator molecules appears to be right-tailed with respect to the number of scaffold levels. The most frequently observed scaffold trees (root level and level 1) were derived in conjunction with Bemis−Murcko scaffolds for both classes of molecules and are shown in Figure 5C−E and Supporting Information Table S1. A cutoff of 10 was used for the frequency. We find that at least 9 most frequent scaffolds exclusively observed in aggregators contain a sulfur atom (Figure 5C), while only 1 most frequent scaffold exclusively observed in nonaggregators contains a sulfur atom (Figure 5D). This result is consistent with the trend in content of sulfur atoms observed in Figure 4A,B.

To compare the similarity and diversity of data sets, the Jaccard/Tanimoto coefficient is one of the metrics commonly used.[38,39] It is defined as the ratio of the intersecting set to the union set, as shown in the following equation

$$T = \frac{N_c}{N_a + N_b - N_c} \tag{12}$$

Where N_a, N_b, and N_c denote the number of elements in data set A, B, and C, respectively.

Thus, a calculation of the Jaccard/Tanimoto coefficient between our current aggregators and nonaggregators data sets yielded a value of 0.31, which suggests an overall good enough

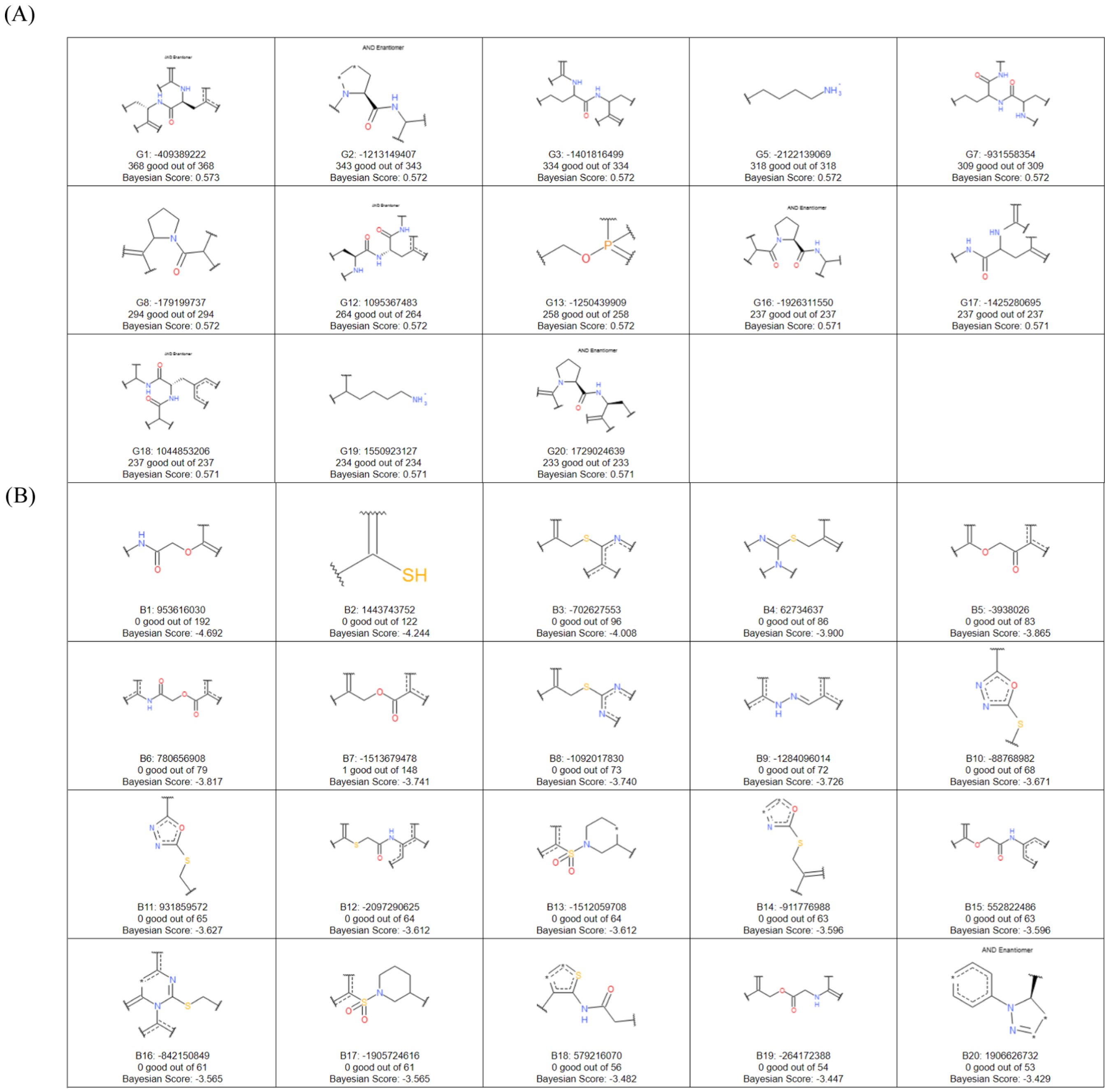


Figure 4. Chemical features are shown with their score, as derived from Naïve Bayesian. (A) Top features favorable for nonaggregation. (B) Top features promoting colloidal aggregation.

dissimilarity between scaffolds promoting colloidal aggregation and those driving nonaggregation.

3.4. Matched Molecular Pair Analysis.

To further decipher the structural context of aggregation-promoting chemical moieties, we have generated matched molecular pair data and output the reaction representing the transformation from fragment A to fragment B in each MMP. We have primarily focused on the case where the matched molecular pair contains a fragment that belongs to a nonaggregator whereas the other fragment belongs to a colloidal aggregator. A detailed list derived from matched molecular pair analysis is shown in Supporting Table S2. Examples of MMPs shown in Figure 6 illustrate that on a phenyl 1,3-thiazole amide scaffold, replacing a pyridine side chain with bromofuran results in turning on aggregation properties (MMP shown in a green rounded rectangle). Likewise, on the benzylamide 1,3,4-thiadiazole scaffold, replacing the hydrogen atom on the thiadiazole ring with furan promotes aggregation (MMP shown in a blue rounded rectangle).

Furthermore, we have analyzed all the changes in aggregation propensity caused by the MMP transformations. All the occurrences of unique transformations from substructure A to substructure B in different pairs of molecules are merged to calculate the percentage of times each transformation increased, decreased, or had no effect on the aggregation. MMP transformations with frequency of occurrence ≥50, and their effect on the aggregation (same

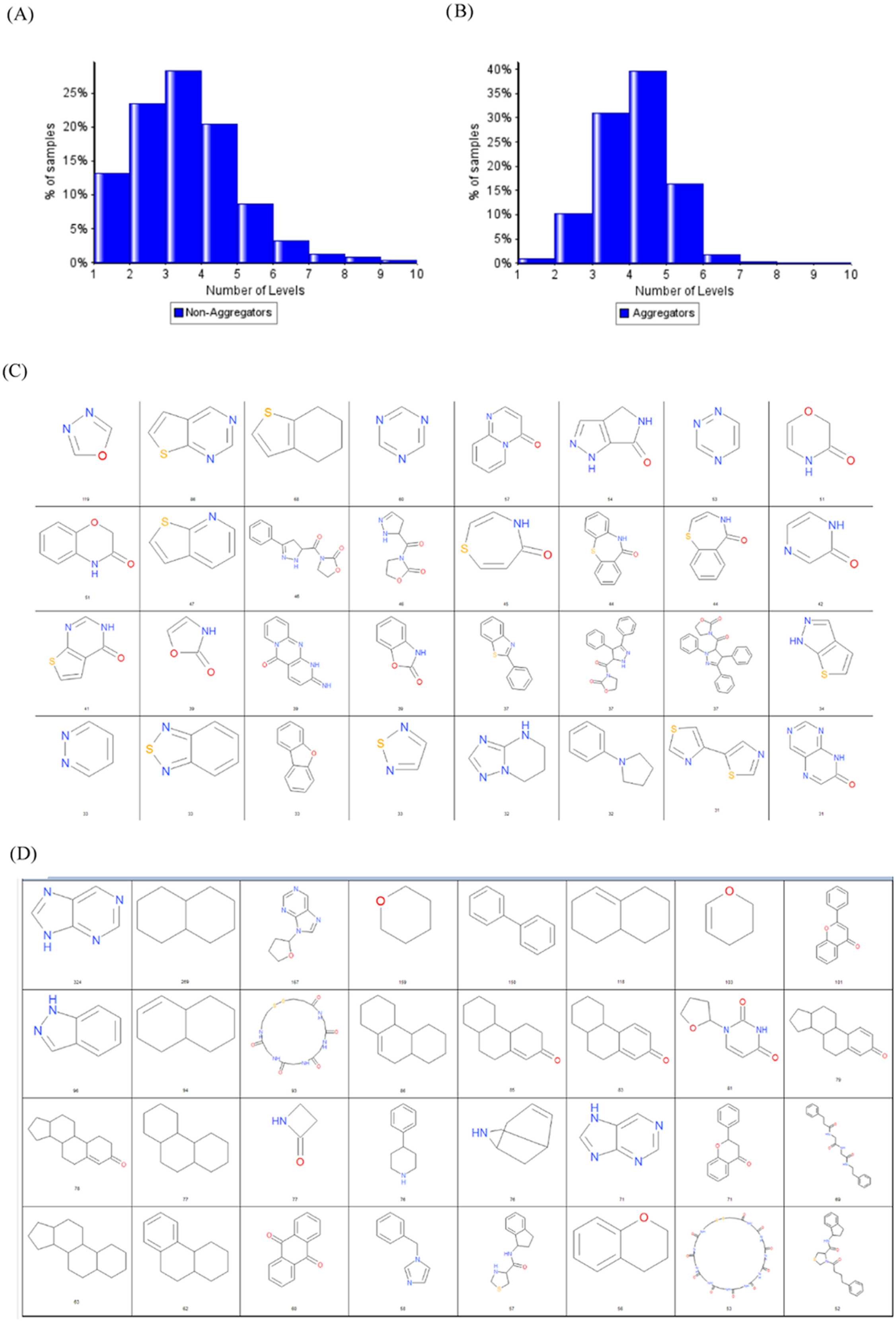


Figure 5. continued

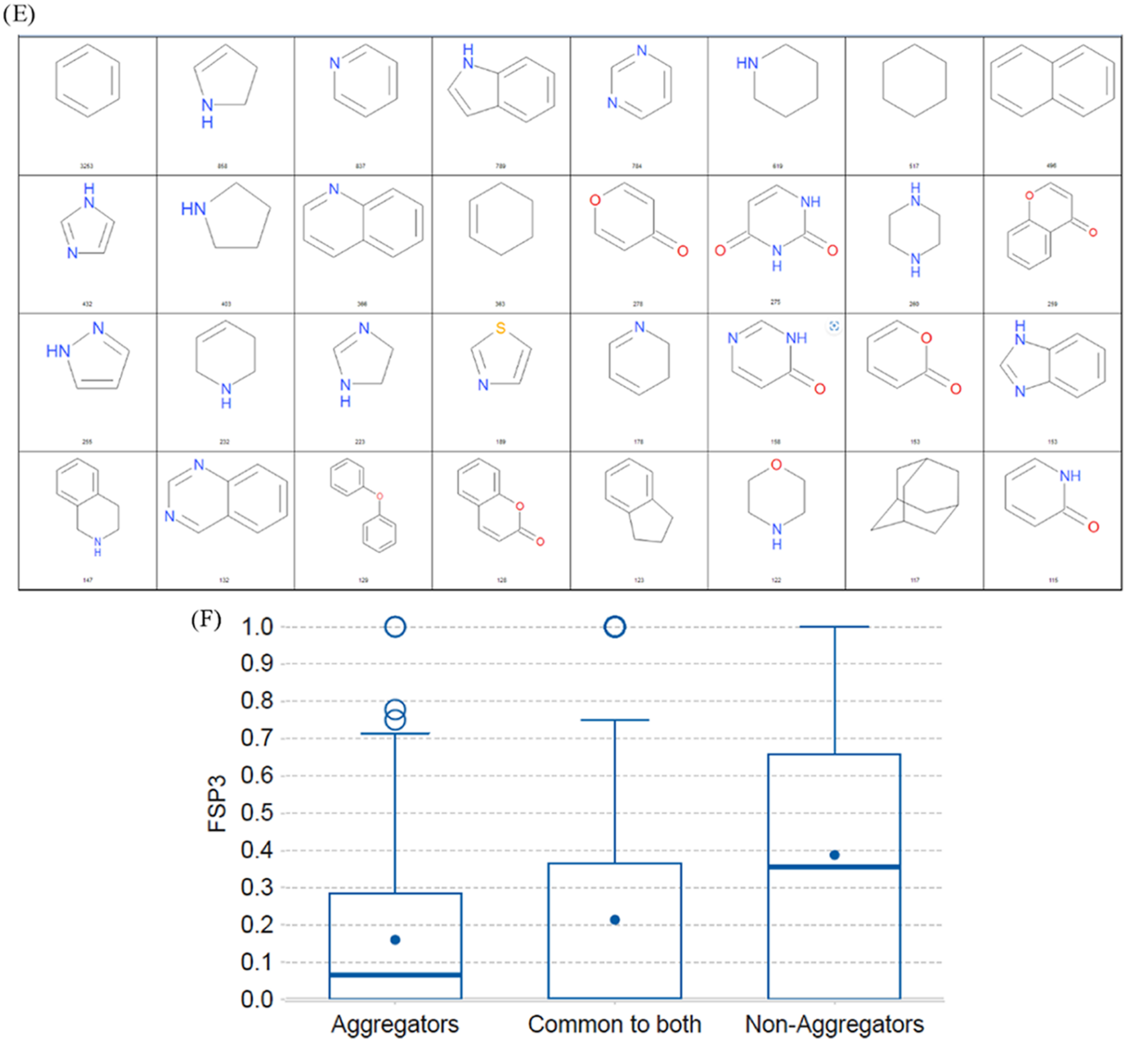


Figure 5. Scaffold Tree and Bemis−Murcko scaffolds derived from the current data set. Panels (A) and (B): Distribution of the number of levels in the scaffold trees for (A) nonaggregators and (B) colloidal aggregators. Panels (C) through (E): Scaffold Tree (root level and level 1) and Bemis−Murcko scaffolds most frequently common to (C) colloidal aggregators only, (D) nonaggregators only, and (E) both nonaggregators and aggregator molecules. (F) Box plot of Fsp3 for scaffolds frequently observed in aggregators only, nonaggregators only, and in both classes of molecules. (D) shows that the scaffolds most frequently present in nonaggregators are sp3 carbon atom-rich molecules. This observation is corroborated by (F), which shows the trend in Fsp3 values for each group of scaffolds. Results are consistent with the increased number of sp3 carbon atoms in the nonaggregator molecules as compared to aggregators, as can also be seen from the bigger size of their filled-cube data points shown in Figure 2, Figure 4A as compared to Figure 4B, and (D) in comparison to (C). Examination of (E) suggests that most used aromatic rings such as phenyl, pyridine, pyrimidine, and indole are frequently found in both classes of molecules. Interestingly, cyclohexyl and piperidine scaffolds, which are commonly used aliphatic bioiosteres for phenyl and pyridine, are also listed among the most frequent scaffolds present in both aggregators and nonaggregators. In our data set, we found 249 scaffolds specific to aggregators, 327 scaffolds specific to nonaggregators, and 135 nonspecific scaffolds (i.e., common to both classes of molecules), with frequency >10.

propensity, increased propensity, or decreased propensity) are shown in Table 5, sorted with respect to percent increase. We have found that while the vast majority of MMP transformations have a neutral effect on aggregation propensity, the following MMP transformations: [H] ≫ [O−], [H] ≫ [N+](=O)[O−], and N ≫ [H], promote colloidal aggregation with a top three increase of 2.8, 2.6, and 1.1%, respectively. Besides, the following MMP transformations: C(=O)C ≫ [H], C ≫ O, and C(=O)[O−] ≫ [H], promote non-aggregation with a top three decrease of 2.5, 1.9, and 1.6%.

3.4.1. Application to Chemical Library Triage for Plate Selection and Purchase. To increase the probability of finding hit molecules in one of our internal drug discovery programs, we have aimed at selecting compound plates with high information content that could be purchased and used in a high throughput screening assay. Therefore, we have utilized predictive models discussed herein to virtually analyze 2 commercially available chemical libraries from a commercial vendor. The content of these libraries and their novelty as compared to our internal Sanofi Corporate Database are described in Table 6. Drug-like property distributions of both libraries are shown in Figure 7A,B.

We have used the deep neural network model to annotate molecules predicted to be aggregators with a probability ≥80%, in addition to using various structural alert filters. Examples include HTS, risky, and reactive filters as described and implemented within Pipeline Pilot, nasty[40] and PAINS[41] filters. Results shown in Table 7 indicate that the model predicts a reasonable propensity for colloidal aggregation of 16 and 28% for the prefiltered and unfiltered libraries, respectively. We

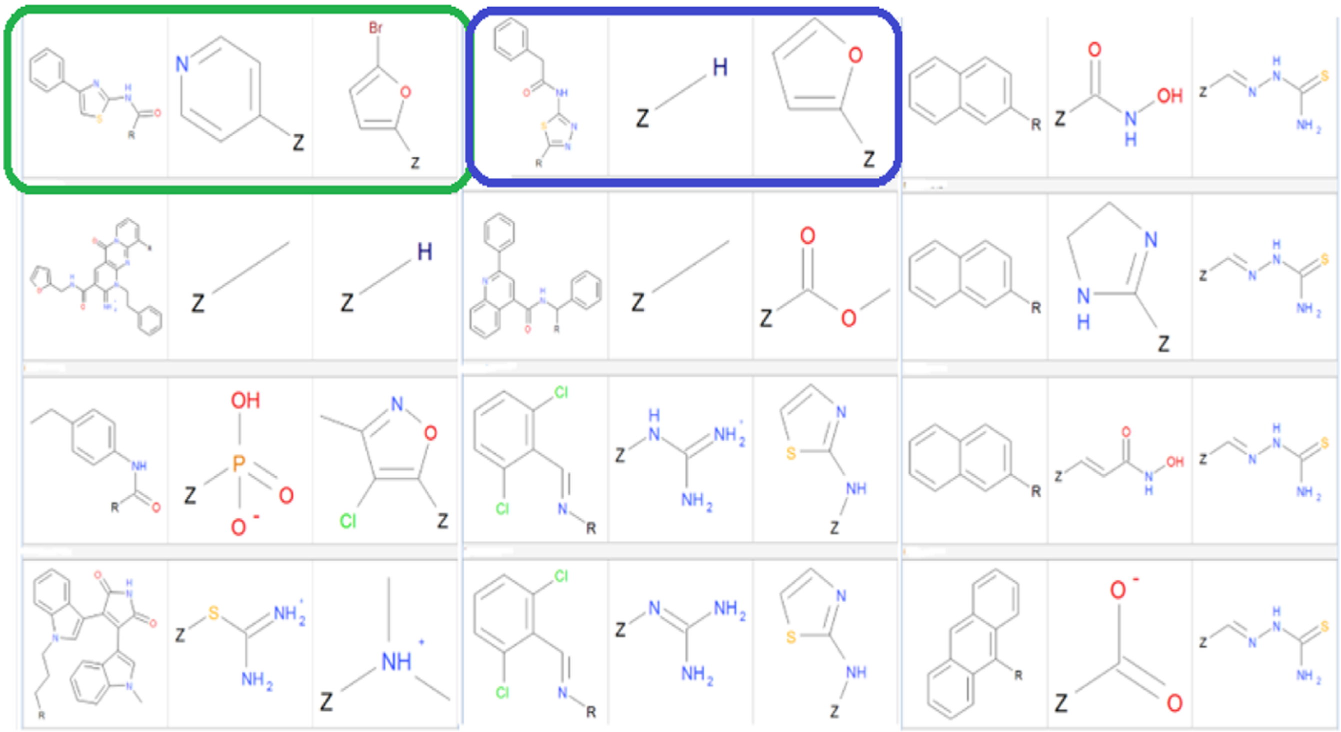


Figure 6. Examples of MMPs that increase the propensity for molecular aggregation. For each of the 12 triplets of fragments, the common core is the leftmost, followed by R groups promoting nonaggregation and those promoting aggregation in the context of the given scaffold, respectively. MMPs involving the phenyl thiazole amide scaffold and the benzylamide thiadiazole core are shown embedded in a green rounded rectangle and a blue rounded rectangle, respectively.

finally selected plates from both libraries, using a multiobjective Pareto optimization[42−44] aimed at minimizing the probability for aggregation and maximizing both molecular diversity and drug-likeness. The latter was calculated using a Bayesian model built using drug-likeness descriptors and molecular fingerprints. The areas under the ROC curves obtained for a training set of 11805 molecules and a test set of 11891 molecules are 0.98 and 0.95, respectively. The corresponding ROC curves are shown in Figure 7C,D.

4. DISCUSSION

Numerous computational studies aimed at predicting colloidal aggregation have been carried out in the last two decades. Previous studies have used Machine Learning techniques including Recursive portioning, Naïve Bayesian, Support Vector Machine, random forest, Extreme Gradient Boosting Decision Trees, K-nearest neighbors, continuous kernel discrimination, principal component analysis, and 2D Tanimoto similarity. Among these techniques, we have chosen to use Naïve Bayesian, random forest, recursive partitioning tree, support vector machine, and PCA techniques. In addition, we explored the use of logistic regression and deep neural networks. We have found that the latter approach and Naïve Bayesian outperformed other techniques by yielding the lowest BER values for the test set (0.10 and 0.08, respectively), as shown in Table 3. We also found that Recursive Partitioning Tree outperforms all of the other models uniquely for the training set (Table 2) but scores worst for both the test and validation sets (Tables 3 and 4). In comparison, the models are less performant on the internal validation set, as shown in Table 4. Encouragingly, we obtained an excellent precision metric on the validation set compared to that derived for the training and test sets (within the range 0.88−0.95). Furthermore, we have investigated for the first time the use of the modelability index as a flag prior to modeling a colloidal aggregation data set. We found that although the data set was assembled from heterogeneous sources, once coupled to the set of judiciously selected descriptors, it yielded excellent values of modelability index for binary classification (up to 0.94, Table 1). Encouragingly, this predicted efficient modelability index was further validated by the summary statistics of the performance metrics (Tables 2 and 3) which showed an accuracy of up to 0.96 and 0.93 for the training and test set, respectively.

To decipher physicochemical properties that drive aggregation, we have utilized most of the molecular descriptors used in previous studies, as a control experiment to validate our models by reproducing known results. We have found that lipophilicity and molecular weight indeed drive small molecule aggregation, as previously shown in the literature.[16] In addition, we have used the following descriptors known to be important for drug developability: Fsp3, number of stereocenters, QED, and molecular complexity. We also included ES_Sum_sOH as a descriptor. Interestingly, our results on important descriptors as derived from random forest (Figure 1A,B) and PCA (Figure 2) reveal that Fsp3 (but not QED) strongly contributes to predicting colloidal aggregation, and independently confirms the importance of previously identified descriptors such as ALogP and molecular weight. That Fsp3 is an important feature for predicting aggregation is consistent with studies by Lovering et al.[20] which showed a strong correlation between this descriptor and experimentally observed solubility, a well-known important property for drug discovery and development. An increase in Fsp3 confers a

Table 5. Changes in Activity Due to MMP Transformations

MP transformation	frequency	percent neutral	percent increase	percent decrease
[H] ≫ [O−]	72	95.8	2.8	1.4
[H] ≫ [N+](=O)[O−]	115	97.4	2.6	0
N ≫ [H]	94	97.9	1.1	1.1
Br ≫ [H]	196	99	0.5	0.5
C ≫ Cl	209	99.5	0.5	0
C ≫ OC	196	99.5	0.5	0
CCC ≫ [H]	186	99.5	0.5	0
O ≫ [H]	452	99.1	0.2	0.7
C ≫ CC	566	99.8	0.2	0
C ≫ [H]	1895	99.6	0.1	0.3
C(=O)C ≫ [H]	80	97.5	0	2.5
OC ≫ [O−]	50	98	0	2
C ≫ O	54	98.1	0	1.9
C(=O)[O−] ≫ [H]	125	98.4	0	1.6
Cl ≫ OC	123	98.4	0	1.6
Br ≫ OC	61	98.4	0	1.6
C#N ≫ [H]	65	98.5	0	1.5
Br ≫ Cl	128	99.2	0	0.8
OC ≫ [H]	535	99.3	0	0.7
CC ≫ [H]	466	99.8	0	0.2
Cl ≫ [H]	463	99.8	0	0.2
F ≫ [H]	402	100	0	0
C ≫ CCC	299	100	0	0
CC ≫ CCC	276	100	0	0
C ≫ F	149	100	0	0
Cl ≫ F	147	100	0	0
O ≫ OC	133	100	0	0
F ≫ OC	123	100	0	0
C ≫ C(C)C	90	100	0	0
C(C)C ≫ CC	85	100	0	0
Br ≫ C	77	100	0	0
C ≫ CO	71	100	0	0
OC ≫ OCC	68	100	0	0
C(C)C ≫ [H]	64	100	0	0
C ≫ C(=O)[O−]	57	100	0	0
Br ≫ F	57	100	0	0
CC(=O)[O−] ≫ [H]	56	100	0	0
C(C)C ≫ CCC	55	100	0	0
CO ≫ [H]	53	100	0	0
c1cccnc1 ≫ c1ccncc1	52	100	0	0
C(=O)[O−] ≫ CC(=O)[O−]	50	100	0	0

Table 6. Vendor Libraries Content

library	total number of samples	total number of plates	number of compounds per plate	novelty with respect to sanofi corporate database (%)
Lib A	234 240	732	320	100
Lib B	298 284	938	267−320	99
target library to purchase	163 200	510	320	100

three-dimensional context to a molecule, thereby allowing it to escape from aromatic "flat land" which is prone to aggregate via $\pi-\pi$ interactions. Figure 8 shows that aggregators exhibit 198 outliers, whereas the nonaggregators exhibit none. It also indicates that overall, aggregators exhibit lower values of Fsp3 than nonaggregators with the median values for aggregators and nonaggregators being 0.18 and 0.36, respectively. This result strikingly agrees with the finding by Life Chemicals according to which the mean Fsp3 increases from 0.36 for 2.2 million molecules in the drug discovery stage and up to 0.47 for 1179 approved drugs.[45] Kombo et al.[25] also reported that ∼84% of marketed drugs had Fsp3 > 0.42. Detailed descriptions of investigations on Fsp3, including chemical strategies explored in order to improve it, have also been reported.[46,47]

Daniel Heller and collaborators have recently shown that aromatic groups and hydrogen-bond acceptors/donors are essential for nanoaggregate formation.[48,49] Our results reported herein are strikingly consistent with their findings since we have found that ES_sum_sOH, Fsp3, the number of aromatic rings, ES_Sum_aaCH (electrotopological state of aromatic carbon atoms), and the number of H-Bond donors are among the most important descriptors which contribute to aggregation, as shown in Figure 1A,B. ES_Sum_sOH in essence combines both the electronegativity character and the topological environment of a single-bonded OH group in a

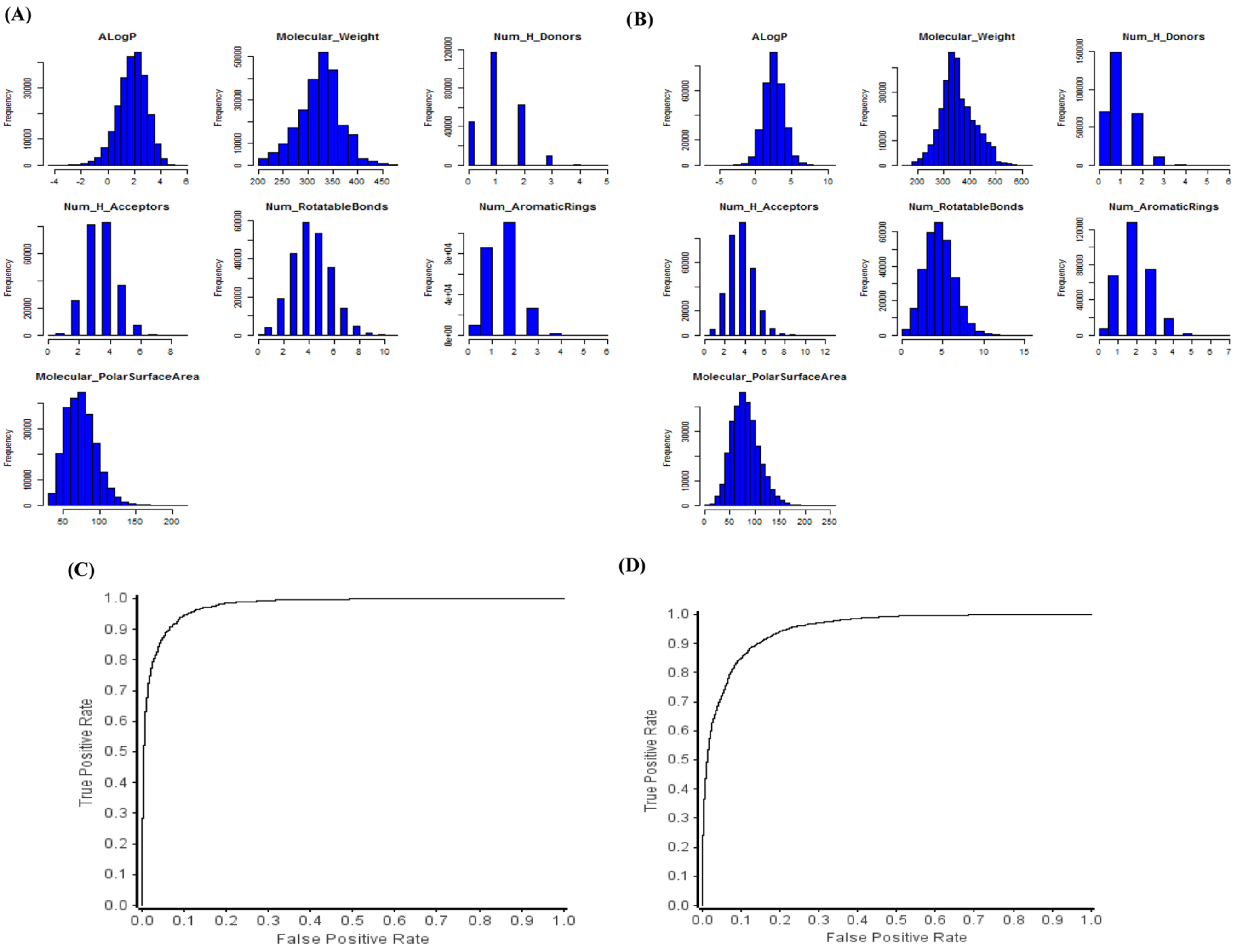


Figure 7. Drug-likeness property distributions and receiver-operated characteristic (ROC) curves. (A) Library A data set property distribution. (B) Library B data set property distribution. (C) ROC curve for the drug-likeness model training set. (D) ROC curve for drug likeness model test set.

Table 7. Summary of Undesirable Chemical Moieties and Aggregation Prediction Results

structural alerts	Lib A	Lib B	total	percent
HTS filters	374	1470	1844	0.35
RISKY	6845	20,630	27,475	5.16
REACTIVE	167	47	214	0.04
NASTY	532	522	1054	0.20
PAINS	895	42	937	0.18
aggregator (prefiltered library)	17,930	18,758	36,688	16.0
aggregator (unfiltered whole library)	6088	140,454	146,542	28.0

molecule.[50] This undoubtedly conveys a physicochemical meaning and interpretation to the "number of OH groups" which was previously identified to be important for predicting aggregation.[15] Our results also further support the findings derived by Shamay et al.[51] from quantitative structure-nanoparticle assembly prediction (QSNAP) studies, which showed that electrotopological molecular descriptors are excellent predictive indicators of the nanoassembly and nanoparticle size. Our finding that the Wiener index, a topological descriptor related to molecular branching, is also of significant importance in contributing to aggregation prediction, as it appears to be consistent with molecular complexity, which is expressed by Fsp3.

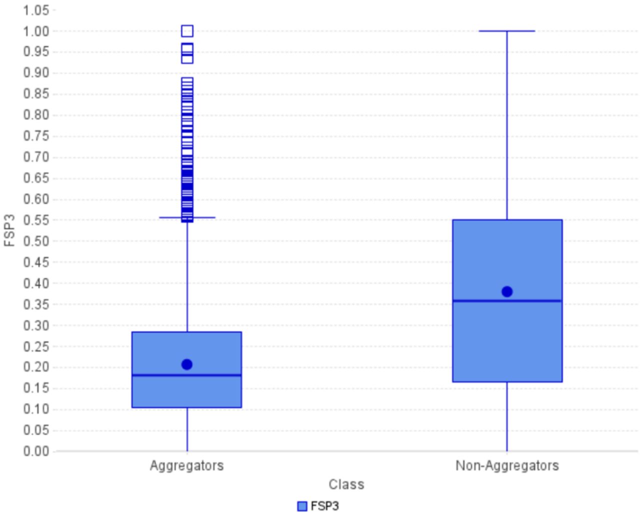


Figure 8. Box plot of Fsp3 for nonaggregators vs aggregators, using the training set.

Moreover, Heller and collaborators[48,49] have demonstrated that nanoaggregate formation can be predicted entirely using drug fragment substructures, which supports our finding that fragments with high propensity to promote aggregation can be clearly distinguished from their counterparts that do not promote aggregation, as shown in Figure 4A,B, 5C,D and Supporting Information Table S1. Interestingly, they have also shown that the colloidal nanoaggregation properties of the phenoxylphenyl scaffold and its phenoxypyridine analogue can be readily turned on and off depending on the incorporation of regioisomeric substitutions. In accordance with their results, we have found that the phenoxylphenyl fragment is frequently observed in both colloidal aggregator and nonaggregator molecules, as shown in Figure 5E (bottom row and third column). We have used Naïve Bayesian to derive the molecular features that drive colloidal aggregation. Furthermore, we have extracted scaffold tree (root and level 1) and Bemis−Murcko scaffolds to better understand the structural difference between aggregators and nonaggregators. Our results have also confirmed previous findings that sulfur-containing five-membered heterocyclic rings appear to promote aggregation as well as uncovered new findings such as the predominant presence of aromatic scaffolds frequently used in medicinal chemistry (and their bioisosteric replacement) in both classes of molecules.

Matched molecular pairs shown in Figure 6 clearly indicate that using the same scaffold, the substituent with more aromatic character tends to promote aggregation. This result agrees with the finding by Chen et al. according to which aromaticity tends to increase nanoaggregation propensity.[48] Furthermore, a survey of all the derived matched molecular pairs is shown both in the main text of this manuscript and in Supporting Information suggests that aggregators are more enriched in hydrogen-bond acceptor/donors pairs and aromatic rings than are their counterparts nonaggregators, as shown in Figure 9A,B. Together with our above-mentioned findings that ES_sum_sOH significantly contributes to colloidal aggregation via the electrostatic contribution of hydrogen-bond donation and that richness in sp3-rich carbon atoms contributes to nonaggregation, our results indicate that intermolecular $\pi-\pi$ interactions and hydrogen-bond interactions are the main drivers of colloidal aggregation. These results agree with recent observations made by Heller and coworkers.[48] Our results on MMPs between aggregators and nonaggregators can be exploited by medicinal chemists in lead optimization of a chemical series whenever a need to rid a compound from colloidal aggregation properties arises.

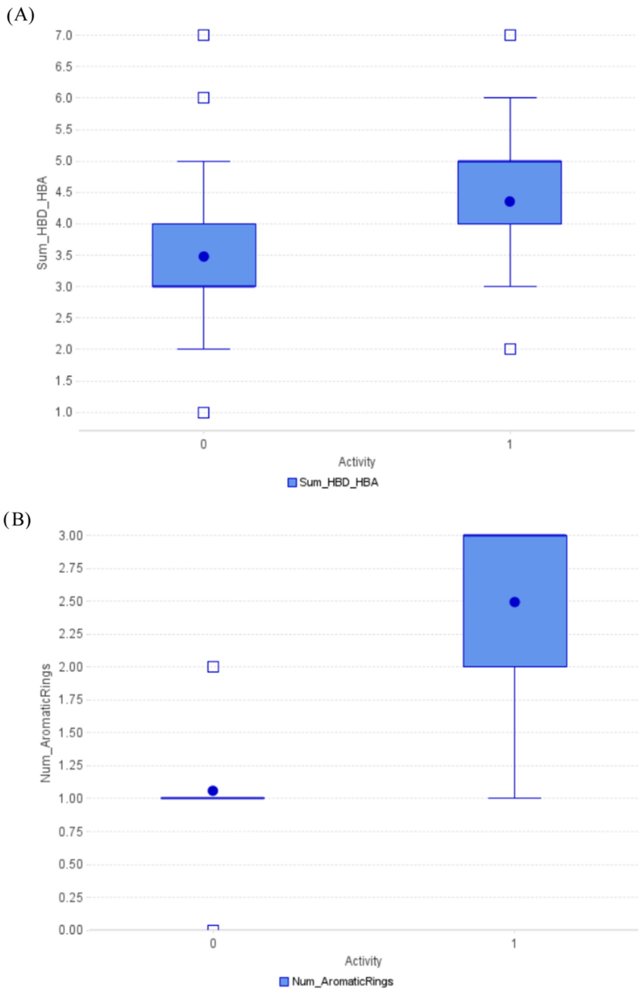


Figure 9. Box plots of molecular properties of MMPs derived between nonaggregators (activity = 0) and colloidal aggregators (activity = 1). (A) Case of the sum of the numbers of hydrogen-bond donors and hydrogen-bond acceptors. (B) Case of the number of aromatic rings.

In a stable colloidal suspension, particles are separated due to repulsive molecular interactions. Particle aggregation can be induced by adding salt to an otherwise stable colloidal suspension. We have found that colloidal aggregation is linked to the presence of aromatic heterocycles mainly involved in π-stacking, hydrogen bonding, and polar interactions. These aromatic heterocyclic compounds can be involved in hydrophobic interactions between themselves and cation-π interactions with the positive charge of the salt. Just like hydrophobic interactions between apolar moieties are promoted in water to maximize stronger water−water hydrogen bonding and polar interactions, it appears that in the presence of a salt, hydrophobic, π-stacking, and hydrogen bonding interactions between substituted aromatic heterocycles are increased, thereby favoring aggregation and maximizing stronger salt-bridge and cation-π interactions involving the salt ions. Whereas electron-donating substituents promote cation- π interactions, electron-withdrawing groups in compounds can change the sign of the aromatic ring quadrupole and promote anion-π interactions between the salt anion and the substituted heterocycle. That the latter aggregation-prone moieties have been found to frequently contain sulfur appears to be consistent with the larger size of the sulfur atom. A detailed matched molecular pairs analysis has suggested that most of the derived transformations (one cut) do not affect the activity, as shown in Table 5. The top-10 most frequent transformations which are 100% neutral in changing the activity are as follows, F ≫ [H], C ≫ CCC, CC ≫ CCC, C ≫ F, Cl ≫ F, O ≫ OC, F ≫ OC, C ≫ C(C)C, C(C)C ≫ CC, and Br ≫ C. In a hit or lead optimization campaign where structural modifications can negatively affect the DMPK properties of a compound of interest, such transformations with 100% neutrality toward aggregation activity can be of significant interest, provided that the biological activity at the protein target is not negatively perturbed. Interestingly, the MMP transformation [H] ≫ [N+](=O)[O−] which has a 2.7% increase in aggregation activity involves a nitro group, a structural alert or a toxicophore commonly avoided during SAR campaign, regardless of its historical use in drugs.[52] Thus, in addition to its well-known property as an agent causing mutagenicity and genotoxicity, a nitro functional group can be an aggregation-promoting agent when attached to a phenyl group, as exemplified in a few transformations (Supporting Information Table S2).

We have prospectively applied our aggregation models in contributing to facilitate a selection of 510 plates rich in diverse drug-like molecules, devoid of undesirable structural alerts, and less likely to aggregate in solution, thereby avoiding compounds prone to be frequent hitters and to interfere with the assay itself. To probe whether the filtering of the 2 commercial libraries by using our aggregation model significantly affects the chemical coverage of the infiltrated libraries, we have quantified the chemical diversity of these libraries prior to and after the filtering process, using their total number of fingerprint features divided by the total number of molecules, as implemented within Pipeline Pilot. We have found that prior to filtering with the structural alerts described in Table 7, Lib B and Lib A consisted of 1.284 and 1.975 fingerprint features per molecule, respectively. After carrying out all the filtering including the use of aggregation model, Lib B and Lib A, resulted in 2.58 and 1.96 fingerprint features per molecule, respectively. This marginal change in the chemical space coverage testifies that these chemical libraries were indeed fairly curated at the design stage. Indeed, this fact agrees with the distributions of drug-like properties for these libraries, as shown in Figure 7A,B. Moreover, that these libraries were full of novel molecules as compared to our internal compound archive and the data set already screened in the project, was pleasing. Finally, some of the aggregation models previously developed and published have been made available to the scientific community via web servers. Even though such models can readily be accessible, their usage by industrial researchers is limited if not precluded due to intellectual property protection issues. This reiterates the need for building internal, proprietary aggregation models.

We have described new modeling and data analysis in this article, previously never described in any computational studies of colloidal aggregation. To build models, we determined and used the following molecular descriptors related to drug developability: Fsp3, QED, number of stereocenters, stereochemical complexity, shape complexity, and synthetic accessibility score. We utilized Deep Neural Network and Logistic Regression as classification algorithms. Furthermore, we have used large-scale automated MMPA[53] to explore structural context-dependence of colloidal aggregation, scaffold trees to describe and analyze scaffolds, and exploited Naïve Bayesian to derive good and bad chemical features akin to aggregation. Whereas previous studies have focused on solely using the Bemis−Murcko scaffold to decipher features important for nonaggregation and colloidal aggregation, we have used both Bemis−Murcko and scaffold tree approaches to also determine molecular features commonly found in both classes of molecules. Furthermore, we illustrated how AI/ML models aimed at predicting colloidal aggregation can be deployed and applied in triaging compounds to be purchased for HTS campaigns in a drug discovery setting. This multifaceted, novel approach toward understanding colloidal aggregation has allowed us to determine important contributors discussed herein, with the fraction of sp3 carbon atoms emerging for the first time as the most important determinant molecular descriptors.

Our computational studies described herein have solely focused on using AI/ML techniques that are based on molecular descriptors due to their inherent ease of interpretability and explainability, which encourages the medicinal chemistry community to prospectively apply the derived models and the SAR lessons learned from them. Alternative methods being intensively developed and used to predict molecular properties and biological activity are well documented under the umbrella of Deep QSAR (for recent and excellent reviews, see Tropsha et al.[54] and Xu[55]). Examples of alternative featurization schemes of molecules, coupled with various deep learning architectures include molecular image recognition-based convoluted neural networks (CNN), molecular graph-based convoluted networks (GCN), smiles strings-based natural language processing (NLP) mechanisms such as recurrent network (RNN) and long short-term memory network (LSTM), etc. In a continuing effort to increase the accuracy of AI/ML models aimed at predicting colloidal aggregation, we are actively investigating some of the above-mentioned algorithms, and our results will be reported elsewhere.

5. CONCLUSIONS

To select small molecules with favorable screening properties, we have used various AI/ML techniques aimed at predicting colloidal aggregation. We have found that derived predictive classification models consistently and successfully discriminate aggregator molecules from their nonaggregator counterparts and outperform previous studies with existing models. In addition to identifying key molecular descriptors that influence aggregation, we have deciphered structural features that (a) influence aggregation in general; by using not only Bemis−Murcko framework analysis, but also scaffold tree and Naïve Bayesian aimed at detecting good and bad features; (b) influence aggregation in specific contexts by carrying out a detailed matched molecular pair analysis; and (c) are common to both aggregators and nonaggregators and thereby rendering difficult the task of predicting and discriminating colloidal aggregators from their nonaggregator counterparts. An analysis of the molecular descriptors responsible for colloidal aggregation suggested that the fraction of sp3 carbon atoms, ES_sum_sOH, $A \log P$, and molecular weight make the strongest impact. Fingerprints-based Naïve Bayesian, scaffold tree analysis, and MMPA have revealed good and bad chemical features/scaffolds contributing the most to colloidal aggregation. Not only does MMPA show that context-dependent substitutions can assist in removing colloidal aggregators but it also suggests molecular transformations that are neutral with respect to promoting aggregation. We have illustrated how the use of AI/ML techniques to predict colloidal aggregation can assist both hit finding and lead optimization efforts in drug discovery and development settings.

■ ASSOCIATED CONTENT

Data Availability Statement

Files for the training set, test set, and examples of Pipeline Pilot protocols used to build predictive models aimed at reproducing data submitted for publication to this journal are provided in the attached file Updated-Zip-File-for-MS.zip. Parameters used for models building are amply listed in Section 2 of the manuscript. Furthermore, the word file named "README_ROADMAP_FILE.docs" which is included in the zip file, provides additional details of the software version, Pipeline Pilot learners and components used in the manuscript, and examples of derived output files.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c02886.

Most frequently observed scaffold trees (root level and level 1) were derived in conjunction with Bemis–Murcko scaffolds for both classes of molecules are shown in Table S1 (XLSX)

A detailed list derived from matched molecular pair analysis is shown in (Table S2) (XLSX)

Training set, test set, output examples, pipeline pilot components and protocols (ZIP)

AUTHOR INFORMATION

Corresponding Author

David C. Kombo − *Integrated Drug Discovery, Sanofi, Cambridge, Massachusetts 02141, United States;* orcid.org/0009-0002-7746-0358; Email: David.Kombo@sanofi.com

Authors

J. David Stepp − *Integrated Drug Discovery, Sanofi, Cambridge, Massachusetts 02141, United States*

Sungtaek Lim − *Integrated Drug Discovery, Sanofi, Cambridge, Massachusetts 02141, United States*

Bettina Elshorst − *CMC Synthetics Early Development Analytics, Sanofi, Frankfurt 65926, Germany*

Yi Li − *Integrated Drug Discovery, Sanofi, Cambridge, Massachusetts 02141, United States*

Laura Cato − *Molecular Oncology, Sanofi, Cambridge, Massachusetts 02141, United States*

Maysoun Shomali − *Molecular Oncology, Sanofi, Cambridge, Massachusetts 02141, United States*

David Fink − *Integrated Drug Discovery, Sanofi, Cambridge, Massachusetts 02141, United States*

Matthew J. LaMarche − *Integrated Drug Discovery, Sanofi, Cambridge, Massachusetts 02141, United States;* orcid.org/0000-0002-6588-2302

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c02886

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank Sanofi U.S. for financial support. We would also like to thank our colleagues from the Sanofi internal review body, especially Christoph Grebner and Sukanthini Thurairatnam for carefully reading the manuscript and providing suggestions.

ABBREVIATIONS USED

AI:artificial intelligence; **ML**:machine learning; **BER**:balanced error rate; **AUC**:area under the curve; **ROC**:receiver operating characteristic curve; **MODI**:modelability index; **BAG**:bootstrap aggregation; **OOB**:out-of-bag; **XGBoost**:extreme gradient boosting; **RF**:random forest; **NB**:Naïve Bayesian; **RP**:recursive partitioning; **SVM**:support vector machine; **KNN**:K-nearest neighbors; **CKD**:continuous kernel discrimination; **TP**:true positive; **TN**:true negative; **FP**:false positive; **FN**:false negative; **GLM**:generalized linear model; **PCA**:principal component analysis; **PAINS**:pan-assay interference compounds; **ES**:electrotopological state indices; **Fsp3**:fraction of sp^3 carbon atoms; **QED**:quantitative estimate of drug-likeness; **ECFP**:extended connectivity functional class molecular fingerprints; **MMPA**:matched molecular pairs analysis; **MMP**:matched molecular pairs; **QSNAP**:quantitative structure-nanoparticle assembly prediction; **HTS**:high throughput screening; **STD**:saturation transfer difference; **NMR**:nuclear magnetic resonance; **TTD**:therapeutic target database

REFERENCES

(1) Duan, D.; Torosyan, H.; Elnatan, D.; McLaughlin, C. K.; Logie, J.; Shoichet, M. S.; Agard, D. A.; Shoichet, B. K. Internal Structure and Preferential Protein Binding of Colloidal Aggregates. *ACS Chem. Biol.* **2017**, *12*, 282−290.

(2) McLaughlin, C. K.; Duan, D.; Ganesh, A. N.; Torosyan, H.; Shoichet, B. K.; Shoichet, M. S. Stable Colloidal Drug Aggregates Catch and Release Active Enzymes. *ACS Chem. Biol.* **2016**, *11*, 992−1000.

(3) Owen, S. C.; Doak, A. K.; Wassam, P.; Shoichet, M. S.; Shoichet, B. K. Colloidal aggregation affects the efficacy of anticancer drugs in cell culture. *ACS Chem. Biol.* **2012**, *7*, 1429−1435.

(4) Dotolo, S.; Marabotti, A.; Facchiano, A.; Tagliaferri, R. A review on drug repurposing applicable to COVID-19. *Brief Bioinform.* **2021**, *22*, 726−741.

(5) Hossain, M. S.; Hami, I.; Sawrav, M. S. S.; Rabbi, M. F.; Saha, O.; Bahadur, N. M.; Rahaman, M. M. Drug Repurposing for Prevention and Treatment of COVID-19: A Clinical Landscape. *Discoveries* **2020**, *8*, No. e121, DOI: 10.15190/d.2020.18.

(6) Singh, T. U.; Parida, S.; Lingaraju, M. C.; Kesavan, M.; Kumar, D.; Singh, R. K. Drug repurposing approach to fight COVID-19. *Pharmacol. Rep.* **2020**, *72*, 1479−1508.

(7) Saxena, A. Drug targets for COVID-19 therapeutics: Ongoing global efforts. *J. Biosci.* **2020**, *45*, No. 87, DOI: 10.1007/s12038-020-00067-w.

(8) Papapetropoulos, A.; Szabo, C. Inventing new therapies without reinventing the wheel: the power of drug repurposing. *Br. J. Pharmacol.* **2018**, *175*, 165−167.

(9) O'Donnell, H. R.; Tummino, T. A.; Bardine, C.; Craik, C. S.; Shoichet, B. K. Colloidal Aggregators in Biochemical SARS-CoV-2 Repurposing Screens. *J. Med. Chem.* **2021**, *64*, 17530−17539.

(10) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J. Med. Chem.* **2003**, *46*, 4477−4486.

(11) *Cerius2*, 4.2th ed.; Accelrys, Inc.: San Diego.

(12) Feng, B. Y.; Shelat, A.; Doman, T. N.; Guy, R. K.; Shoichet, B. K. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **2005**, *1*, 146−148.

(13) Rao, H.; Li, Z.; Li, X.; Ma, X.; Ung, C.; Li, H.; Liu, X.; Chen, Y. Identification of small molecule aggregators from large compound libraries by support vector machines. *J. Comput. Chem.* **2010**, *31*, 752−763.

(14) Irwin, J. J.; Duan, D.; Torosyan, H.; Doak, A. K.; Ziebart, K. T.; Sterling, T.; Tumanian, G.; Shoichet, B. K. An Aggregation Advisor for Ligand Discovery. *J. Med. Chem.* **2015**, *58*, 7076−7087.

(15) Yang, Z. Y.; Yang, Z. J.; Dong, J.; Wang, L. L.; Zhang, L. X.; Ding, J. J.; Ding, X. Q.; Lu, A. P.; Hou, T. J.; Cao, D. S. Structural Analysis and Identification of Colloidal Aggregators in Drug Discovery. *J. Chem. Inf. Model.* **2019**, *59*, 3714−3726.

(16) Reker, D.; Bernardes, G. J. L.; Rodrigues, T. Computational advances in combating colloidal aggregation in drug discovery. *Nat. Chem.* **2019**, *11*, 402−418.

(17) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv Rev.* **2001**, *46*, 3−26.

(18) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral

bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.

(19) Ritchie, T. J.; Macdonald, S. J. The impact of aromatic ring count on compound developability–are too many aromatic rings a liability in drug design? *Drug Discovery Today* **2009**, *14*, 1011−1020, DOI: 10.1016/j.drudis.2009.07.014.

(20) Lovering, F.; Bikker, J.; Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **2009**, *52*, 6752−6756.

(21) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90−98.

(22) Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18787−18792.

(23) Clemons, P. A.; Wilson, J. A.; Dančík, V.; Muller, S.; Carrinski, H. A.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 6817−6822.

(24) Méndez-Lucio, O.; Medina-Franco, J. L. The many roles of molecular complexity in drug discovery. *Drug Discovery Today* **2017**, *22*, 120−126, DOI: 10.1016/j.drudis.2016.08.009.

(25) Kombo, D. C.; Tallapragada, K.; Jain, R.; Chewning, J.; Mazurov, A. A.; Speake, J. D.; Hauser, T. A.; Toler, S. 3D molecular descriptors important for clinical success. *J. Chem. Inf Model.* **2013**, *53*, 327−342.

(26) *Pipeline Pilot*, version 21.2.0.2574; Dassault Systems Biovia Corporation: San Diego, CA, USA, 2020.

(27) Irwin, J. J.; Shoichet, B. K. ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(28) https://drugcentral.org.,06/10/2024.

(29) Li, Y. H.; Yu, C. Y.; Li, X. X.; Zhang, P.; Tang, J.; Yang, Q.; Fu, T.; Zhang, X.; Cui, X.; Tu, G.; Zhang, Y.; Li, S.; Yang, F.; Sun, Q.; Qin, C.; Zeng, X.; Chen, Z.; Chen, Y. Z.; Zhu, F. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.* **2018**, *46* (D1), D1121−D1127.

(30) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, No. 8, DOI: 10.1186/1758-2946-1-8.

(31) Dalvit, C.; Caronni, D.; Mongelli, N.; Veronesi, M.; Vulpetti, A. NMR-based quality control approach for the identification of false positives and false negatives in high throughput screening. *Curr. Drug Discovery Technol.* **2006**, *3*, 115−124.

(32) Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data set modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54*, 1−4.

(33) https://mcule.com/.06/10/2024.

(34) https://www.emolecules.com/,06/10/2024.

(35) https://chembridge.com/,06/10/2024.

(36) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(37) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The scaffold tree–visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, *47*, 47−58.

(38) Jaccard, P. The distribution of the flora in the alpine zone. *New Phytol.* **1912**, *11*, 37−50.

(39) Tanimoto, T. An Elementary Mathematical Theory of Classification and Prediction. In *Technical Report*; International Business Machines Corporation, 1958.

(40) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897−902.

(41) Baell, J. B.; Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.

(42) Deb, K.; Agrawal, S.; Pratap, A.; Meyarivan, T. A. A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimization: NSGA-II. In *Parallel Problem Solving from Nature PPSN VI*; Schoenauer, M.et al., Ed.; Lecture Notes in Computer Science; Springer: Berlin, Heidelberg, 2000; p 1917.

(43) http://www-fp.mcs.anl.gov/otc/Guide/OptWeb/multiobj/,06/10/2024.

(44) http://en.wikipedia.org/wiki/Pareto_efficiency,06/10/2024.

(45) https://lifechemicals.com/screening-libraries/fragment-libraries,06/10/2024.

(46) Wei, W.; Cherukupalli, S.; Jing, L.; Liu, X.; Zhan, P. Fsp3: A new parameter for drug-likeness. *Drug Discovery Today* **2020**, *25*, 1839−1845.

(47) Caplin, M. J.; Foley, D. J. Emergent synthetic methods for the modular advancement of sp3-rich fragments. *Chem. Sci.* **2021**, *12*, 4646−4660.

(48) Chen, C.; Wu, Y.; Wang, S. T.; Berisha, N.; Manzari, M. T.; Vogt, K.; Gang, O.; Heller, D. A. Fragment-based drug nano-aggregation reveals drivers of self-assembly. *Nat. Commun.* **2023**, *14*, No. 8340, DOI: 10.1038/s41467-023-43560-0.

(49) https://practicalfragments.blogspot.com/2024/01/what-makes-molecules-aggregate.html?m=1,06/10/2024.

(50) Hall, L. H.; Kier, L. B. The E-State as the Basis for Molecular Structure Space Definition and Structure Similarity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 784−791.

(51) Shamay, Y.; Shah, J.; Işık, M.; Mizrachi, A.; Leibold, J.; Tschaharganeh, D. F.; Roxbury, D.; Budhathoki-Uprety, J.; Nawaly, K.; Sugarman, J. L.; Baut, E.; Neiman, M. R.; Dacek, M.; Ganesh, K. S.; Johnson, D. C.; Sridharan, R.; Chu, K. L.; Rajasekhar, V. K.; Lowe, S. W.; Chodera, J. D.; Heller, D. A. Quantitative self-assembly prediction yields targeted nanomedicines. *Nat. Mater.* **2018**, *17*, 361−368.

(52) Nepali, K.; Lee, H. Y.; Liou, J. P. Nitro-Group-Containing Drugs. *J. Med. Chem.* **2019**, *62*, 2851−2893.

(53) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.

(54) Tropsha, A.; Isayev, O.; Varnek, A.; et al. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nat. Rev. Drug Discovery* **2024**, *23*, 141−155, DOI: 10.1038/s41573-023-00832-0.

(55) Xu, Y. Deep Neural Networks for QSAR. *Methods Mol. Biol.* **2022**, *2390*, 233−260.